

# Business Case: Netflix - Data Exploration and Visualisation

## Table of Contents

### **1. Introduction**

- a. Overview of Netflix<sup>1</sup>
- b. Dataset Description

### **2. Business problem**

- a. Statement of Problem
- b. Objectives

### **3. Data Exploration**

- a. Initial Data Exploration
- b. Non-Graphical Analysis

### **4. Data Pre-Processing**

- a. Unesting
- b. Data Imputation and handling duplicates

### **5. Visual Analysis**

- a. Univariate Analysis
- b. Bi-Variate Analysis

### **6. Insights from Analysis**


- a. Observations on Attributes
- b. Distribution Comments
- c. Univariate/Bivariate Plot Comments

### **7. Business Insights**

- a. Patterns Observed
- b. Implications

### **8. Recommendations and Conclusion**

---

<sup>1</sup>  Netflix\_Buisness\_CaseStudy.ipynb

## Introduction

### Overview of Netflix

Netflix, founded in 1997, has transformed from a DVD rental service into a global streaming giant with over 230 million subscribers. It offers a vast library of more than 10,000 movies and TV shows, including critically acclaimed original content like *Stranger Things* and *The Crown*. As a market leader, Netflix sets the standard for content delivery and user experience, influencing how audiences consume entertainment. Its data-driven approach to content creation and marketing helps the platform understand viewer preferences, ensuring continued relevance in an ever-evolving media landscape. By pioneering the streaming model, Netflix has disrupted traditional media, inspiring a wave of competition and innovation across the industry.

### Dataset Description

The dataset comprises comprehensive listings of movies and TV shows available on Netflix, encapsulating vital information that aids in understanding content distribution and audience preferences. It contains the following key attributes:

- **show\_id**: A unique identifier for each movie or TV show.
- **type**: Indicates whether the content is a movie or a TV show.
- **title**: The title of the movie or TV show.
- **director**: The director responsible for the production.
- **cast**: The actors featured in the movie or show.
- **country**: The country of production, providing insight into regional content offerings.
- **date\_added**: The date when the content was added to Netflix, useful for trend analysis over time.
- **release\_year**: The year the content was originally released, allowing for historical comparisons.
- **rating**: The content's rating, which informs on suitability for different audiences.
- **duration**: The total viewing time in minutes (for movies) or the number of episodes (for TV shows).
- **listed\_in**: The genre classification, which helps analyze genre preferences.
- **description**: A brief summary of the content, useful for understanding themes and narratives.

## **Business problem**

### **Statement of Problem**

The rapid growth of the streaming industry presents both opportunities and challenges for Netflix in optimizing its content strategy. As competition increases, understanding which types of shows and movies resonate with diverse audiences across different regions becomes crucial. This analysis aims to identify trends in content availability, genre preferences, and viewer engagement by leveraging a comprehensive dataset of Netflix's offerings. By answering key questions regarding content distribution, seasonal performance, and the influence of directors and actors, this study seeks to provide data-driven insights that will aid Netflix in making informed decisions about future content production and marketing strategies, ultimately enhancing user satisfaction and subscription growth.

### **Objectives**

1. **Analyze Content Distribution:** Assess the availability of shows and movies across countries to identify regional preferences.
2. **Evaluate Trends:** Examine historical data on movie and TV show releases over the past few decades.
3. **Compare Content Focus:** Investigate the balance between TV shows and movies on the platform.
4. **Determine Optimal Launch Times:** Identify the best periods to release TV shows for maximum viewership.
5. **Assess Actor and Director Impact:** Analyze the influence of directors and actors on the success of various content types.

## Data Exploration

### Initial Data Exploration

1. **Shape of Data:** The dataset contains a total of 8,807 rows and 12 columns, indicating a substantial amount of information about various shows and movies on Netflix.
2. **Data Types:** The dataset comprises the following data types for each attribute: `show_id` (object), `type` (object), `title` (object), `director` (object), `cast` (object), `country` (object), `date_added` (object), `release_year` (int64), `rating` (object), `duration` (object), `listed_in` (object), and `description` (object).
3. **Missing Value Detection :** The dataset exhibits some notable missing values that could affect the analysis. Specifically, the **director** column has **50,643 NaN values**, accounting for **25.07%** of the dataset, which significantly limits insights related to directorial influence on content. Additionally, the **country** column has **11,897 NaN values** (5.89%), indicating missing production country information crucial for geographical analysis. The **cast** column also has **2,146 NaN values** (1.06%), which may impact insights regarding actor contributions. Other columns, such as **rating** with **67 NaN values** (0.03%) and **date\_added** with **158 NaN values** (0.08%), contain relatively low percentages of missing data. Columns like **show\_id**, **listed\_in**, **type**, **title**, **release\_year**, **description**, and both **duration** fields have no missing values, ensuring data integrity in these areas. Addressing the significant missing values in director, country, and cast information will be essential for deriving comprehensive insights from the dataset.

### Non-Graphical Analysis

The dataset reveals a balanced distribution between Movies and TV Shows, with Movies slightly outnumbering TV Shows. The `listed_in` column highlights a wide variety of genres, indicating diverse content offerings on Netflix. There is content produced across multiple countries, with a few countries like the US, India, UK etc.

## Data Pre-Processing

### Unesting

To handle the nested columns in the dataset, such as cast, listed\_in, director, and country, the process of unesting was applied. The unesting function splits values in columns where multiple items are present (e.g., multiple cast members or genres listed together) and explodes them into separate rows. This ensures each unique value, such as each cast member or director, is associated with the corresponding show\_id in individual rows.

Here is the step-by-step approach:

1. **Unesting Columns:** The function `unest_column` splits and explodes columns like cast, listed\_in, director, and country, creating separate rows for each unique value, ensuring there are no nested lists of values within the same row.
2. **Merging DataFrames:** After unesting the relevant columns, the resulting DataFrames were merged back together using show\_id as the key. This ensures that all relevant columns, including the unnested ones, are aligned based on the show or movie.
3. **Handling Leftover Columns:** A separate DataFrame was created for non-nested columns such as type, title, release\_year, and rating, which did not require unesting.
4. **Final Merging:** The unnested DataFrames and the leftover columns DataFrame were merged into one final DataFrame, ensuring all columns and data are in a clean, flat structure for further analysis.

This process allows for a more granular exploration of the dataset, enabling deeper insights into relationships between individual elements like actors, directors, and genres.

## Data Imputation and handling duplicates

In the process of cleaning and preparing the Netflix dataset for analysis, data imputation was a critical step to handle missing values and ensure consistency across the dataset. Here's an overview of the key steps taken for imputing missing values:

### 1. Imputation of Categorical Columns

- **Director:** Since 25% of the data in the 'director' column was missing, we replaced the NaN values with "Unknown", as using the mode would not be appropriate due to the high number of missing entries.
- **Cast:** With around 1% missing values, we used the mode within the same country to fill missing values in the 'cast' column. This ensured that missing cast entries were imputed contextually within a country.
- **Country:** For missing values in the 'country' column, the mode was calculated under each specific director. If the director was unknown, the country remained unchanged.

### 2. Numerical Column Imputation

- **Duration Magnitude:** Missing values in the 'duration\_magnitude' column were filled with the mean of the particular type (i.e., movies or TV shows). This method ensured that average values were used based on the content type.
- **Rating:** For missing values in the 'rating' column, we used the mode within the same director group to fill in the missing entries, as the rating is often consistent across the same director's content.

### 3. Handling Duration Type

Based on the content type, the 'duration\_type' column was imputed. Movies were assigned 'min' (minutes), while TV shows were assigned 'Seasons' to maintain consistency.

### 4. Handling Duplicates

Before proceeding with imputation, we ensured there were no duplicates in the dataset by performing a check on the 'show\_id' column. Duplicates, if any, were dropped to maintain the integrity of the analysis.

This systematic imputation process ensured that missing data was appropriately filled without introducing bias, allowing for accurate and reliable insights from the dataset.

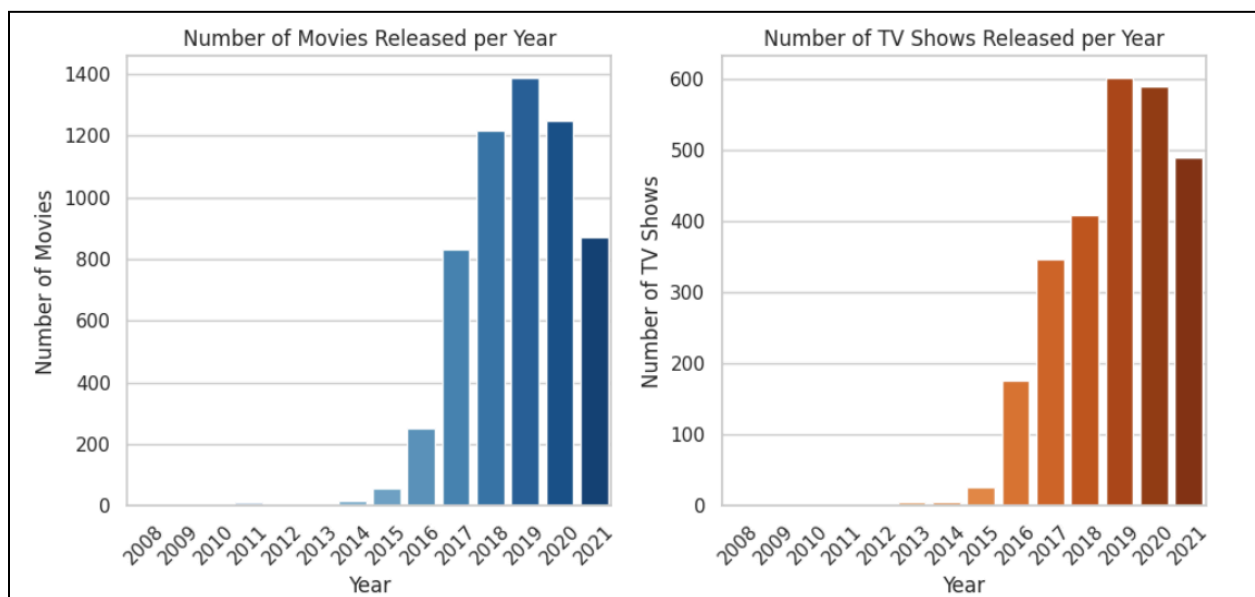
## Visual Analysis

### Univariate Analysis

How has the number of movies released per year changed over the last 20-30 years?

The analysis of the Netflix dataset reveals interesting trends in the release of movies and TV shows over the past 20 years. Starting around 2011, the number of both movies and TV shows released on Netflix saw a steady increase, with a significant surge starting in 2017.

1. **Peak Release in 2019:** The highest number of movies and TV shows were released in 2019, with 1,389 movies and 602 TV shows. This marked Netflix's aggressive content expansion.
2. **Impact of COVID-19:** After the peak in 2019, a noticeable decline occurred in 2020 and 2021, with 2020 seeing 1,249 movies and 590 TV shows released. The drop can largely be attributed to the COVID-19 pandemic and its resulting global production slowdowns and lockdowns.
3. **General Trend:** Overall, the number of releases increased consistently from 2011 through 2019, indicating Netflix's strategy of rapidly expanding its content library to cater to a global audience.

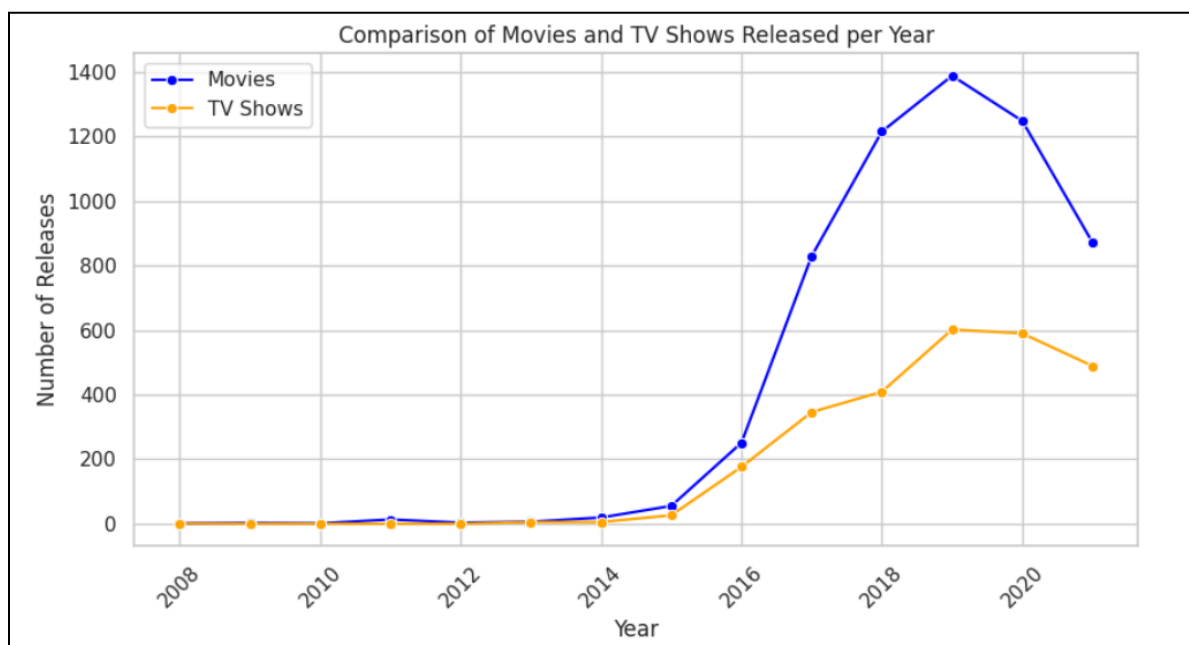


## Comparison of tv shows vs. movies

When comparing the number of movies and TV shows released on Netflix over the years, a few key patterns emerge:

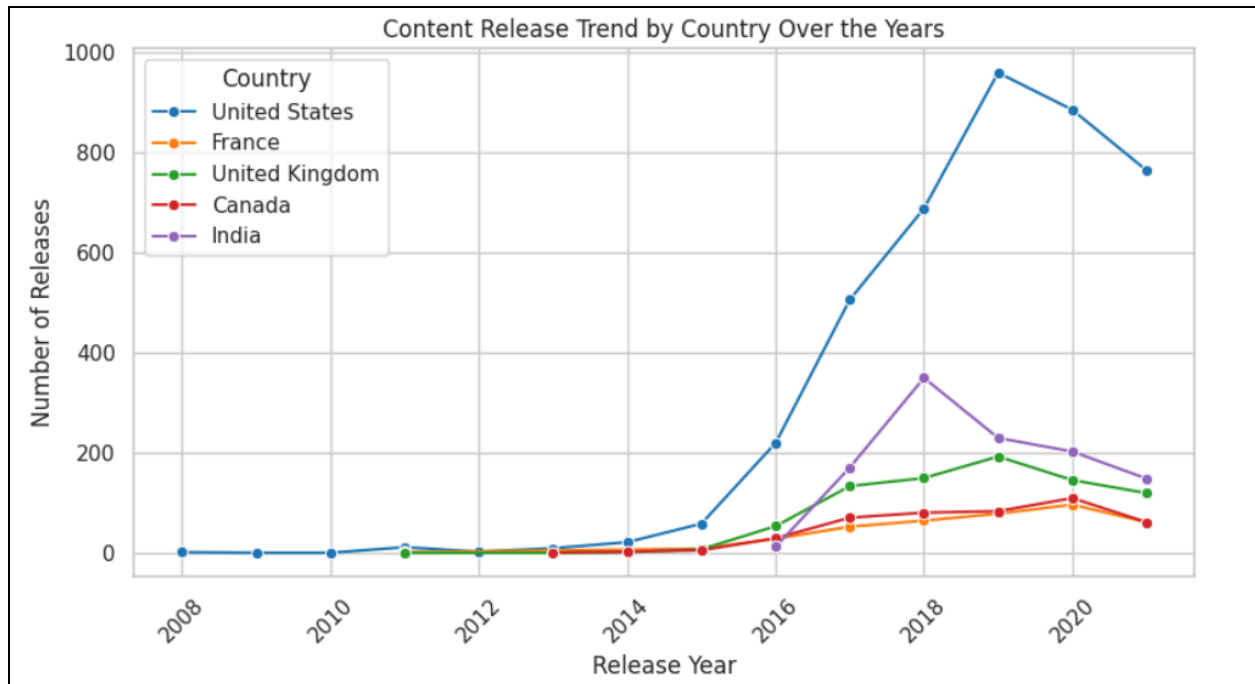
- **Steady Growth in Both Categories:** Between 2011 and 2019, both movies and TV shows saw a consistent increase in releases. However, movies have generally outnumbered TV shows across the years, reflecting Netflix's initial focus on expanding its movie library.
- **2019 Peak for Both:** The highest number of releases for both movies and TV shows occurred in 2019, with a significant lead in movie releases. This coincides with Netflix's aggressive expansion strategy to increase its global content library.
- **Impact of COVID-19:** Post-2019, a decline was observed in both categories due to the COVID-19 pandemic, affecting content production worldwide.

This comparison demonstrates Netflix's strategy of balancing movies and TV shows while slightly favoring the movie category during their content expansion phase.

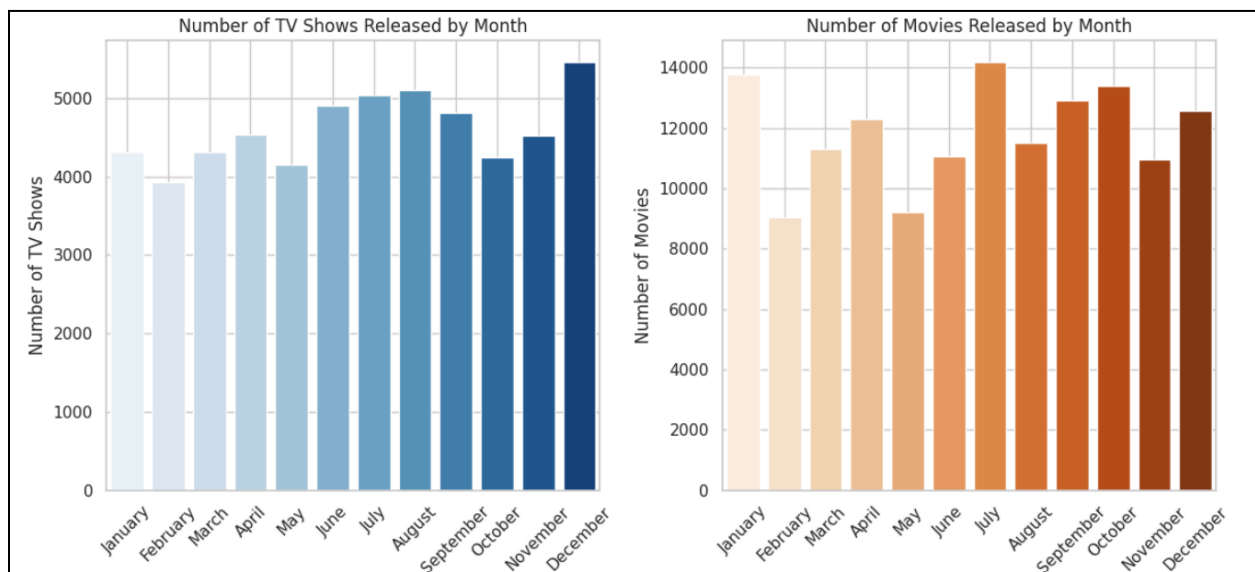


In the early years of Netflix's content production, movies and TV shows were released in nearly equal numbers, particularly until 2014. However, after 2014, a noticeable shift occurred where movies began to dominate in terms of the number of releases. This trend highlights Netflix's increasing focus on producing or acquiring more movies compared to TV shows, particularly in the years leading up to 2019.





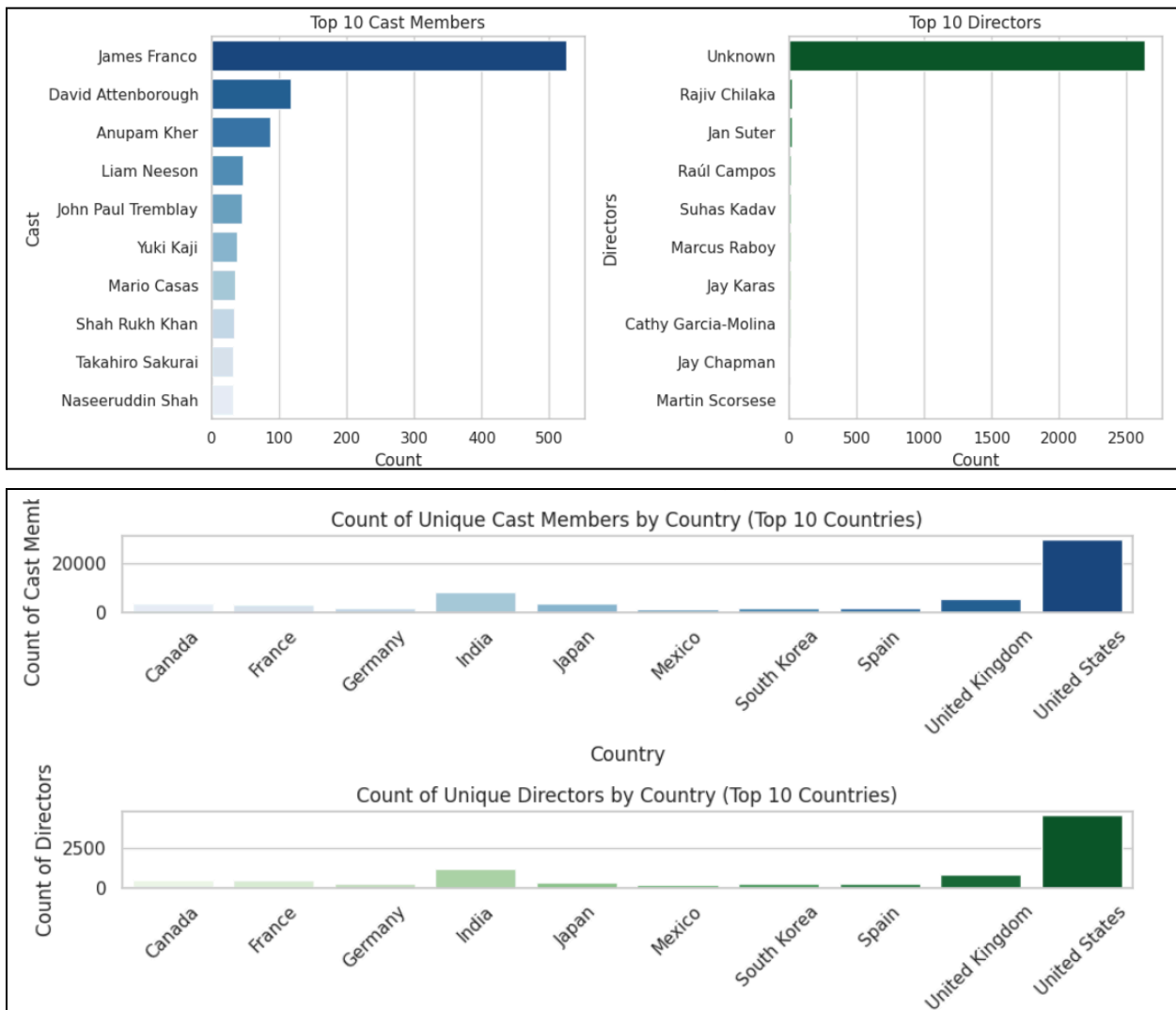
## What is the best time to launch a TV show?



Based on the data analysis, the optimal time to launch movies on Netflix appears to be in July and December, as these months see the highest number of movie releases. This suggests that Netflix strategically releases more content during mid-year and the holiday season to capitalize on increased viewership during vacations and festive periods. For TV shows, December stands out as the best month for releases, likely due to the holiday season, when viewers have more time to binge-watch series.

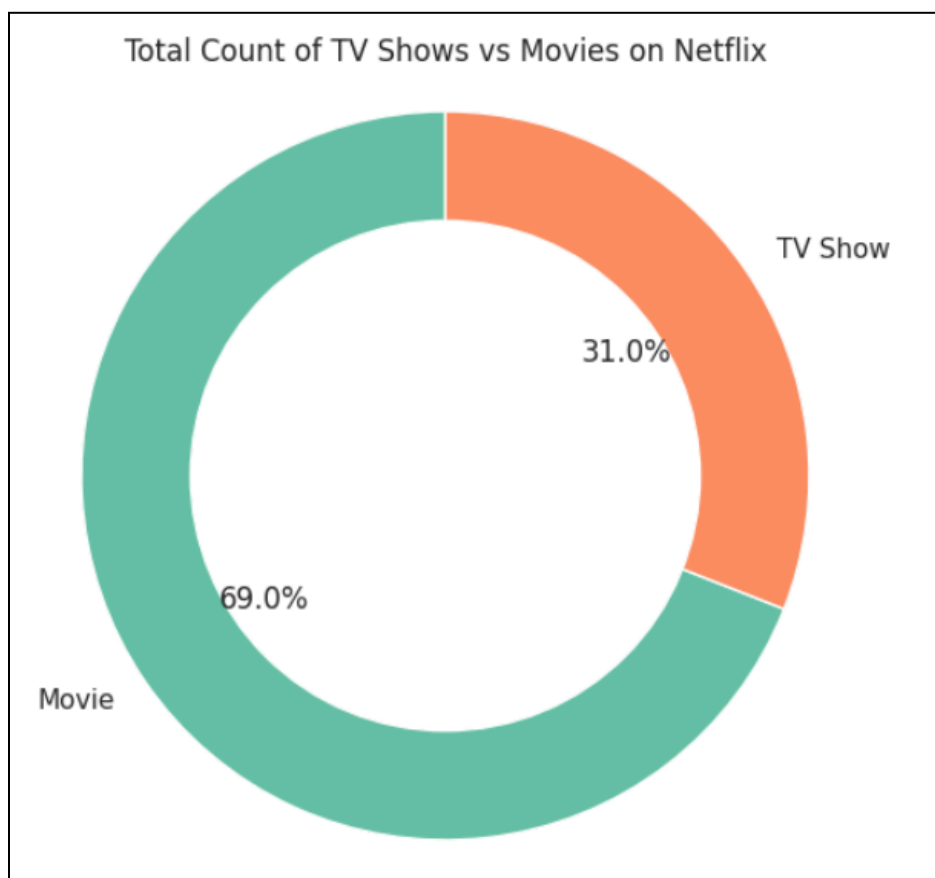
## Analysis of actors/directors of different types of shows/movies

The analysis of actors and directors involved in Netflix shows and movies reveals key insights into the most frequently cast individuals and leading directors. By removing duplicate show and cast combinations, we identified the top 10 cast members with the highest number of appearances across Netflix titles. Similarly, the top 10 directors were determined by counting unique show and director combinations.



## Does Netflix has more focus on TV Shows than movies in recent years

Yes, Netflix has shown a stronger focus on TV shows than movies in recent years. With a total of **5,907 movies** and **2,649 TV shows** in its library, the data indicates that Netflix prioritizes producing and acquiring a broader range of TV content. This trend reflects the growing popularity of binge-watching and serialized storytelling, which TV shows offer. Additionally, the availability of diverse genres and formats in the realm of TV series allows Netflix to cater to varied audience preferences, solidifying its position as a leader in the streaming industry.



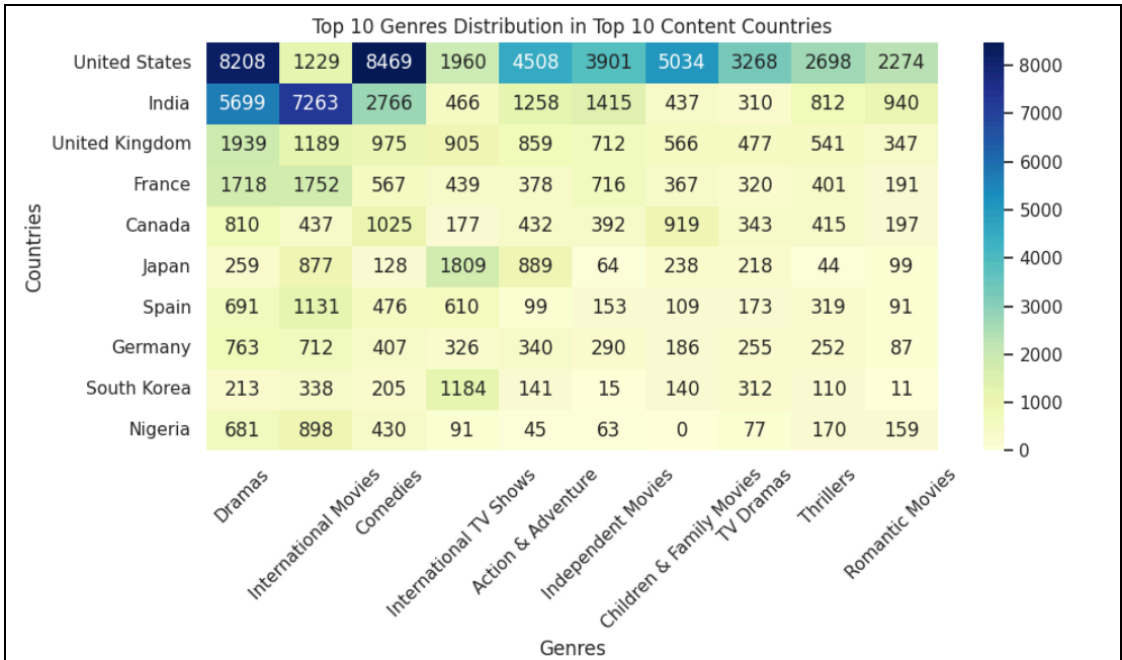
Bi-Variate Analysis

Does Netflix has more focus on TV Shows than movies in recent years

The analysis of Netflix content reveals the distribution of various genres across different countries, highlighting significant trends. Among the top genres, "Dramas" lead the pack with 29,228 entries, followed closely by "International Movies" at 27,244 and "Comedies" at 20,408. The examination further focuses on the top 10 countries contributing to these genres.

By filtering the dataset for these prominent genres, we identified the leading content-producing nations. The resulting heat map illustrates how countries like the United States and India dominate the landscape, particularly in categories such as Dramas, Comedies, and International Movies. This insight emphasizes the content preferences in these regions, where diverse storytelling thrives.

Overall, the heatmap serves as a powerful visualization tool, summarizing the intricate relationships between genre popularity and country-specific content availability on Netflix. This analysis not only aids in understanding current trends but also assists in tailoring recommendations based on regional viewing habits and genre preferences.

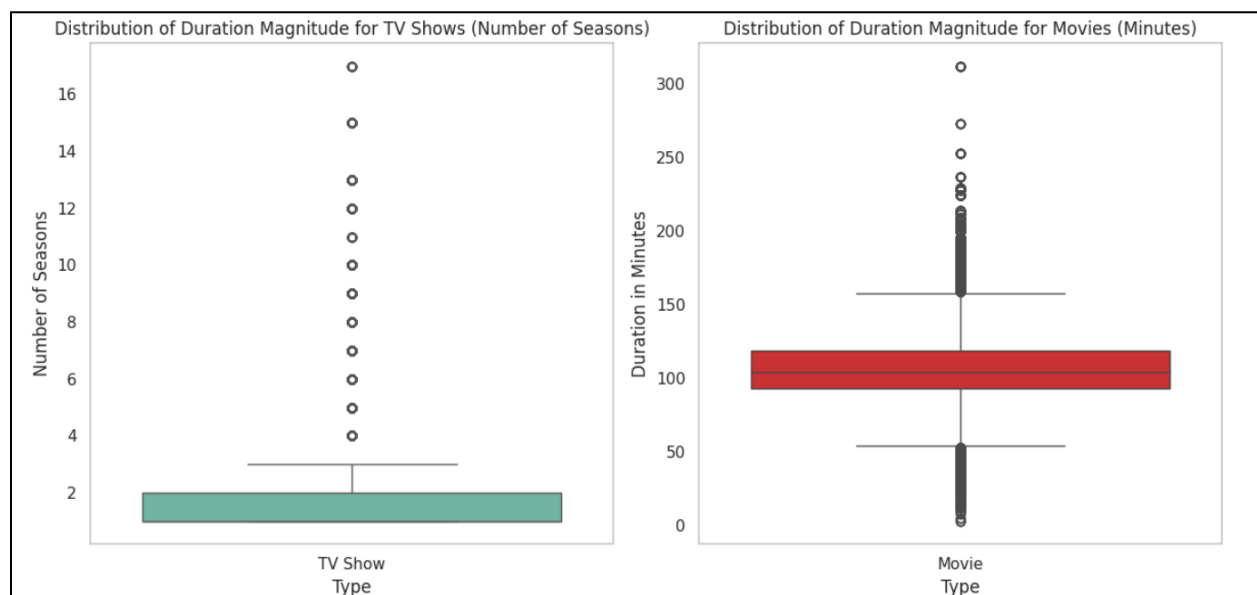


## Understanding what content is available in different countries

This analysis explores Netflix's content by examining TV show durations in episodes and movie lengths in minutes across various countries.

The first boxplot illustrates the distribution of TV show durations, revealing the variability in the number of episodes per series. The second boxplot focuses on movie lengths, highlighting the differences in duration preferences among countries.

Together, these visualizations provide insights into the characteristics of content available in different regions, emphasizing the contrast between TV shows and movies on the platform.



## **Insights from Analysis**

### **Observations on Attributes**

The dataset reveals significant insights into the attributes of Netflix content. The most prevalent genres include Dramas, International Movies, and Comedies, indicating a strong focus on diverse storytelling. The analysis also shows that the highest director count is categorized as "Unknown," likely due to approximately 25% of entries in the director column being missing. This presents an opportunity for recommendations and further exploration of content based on available data. The geographical distribution of cast and director counts highlights that the USA leads, followed by India and the UK, suggesting regional trends in content creation.

### **Distribution Comments**

The distribution of content types indicates a notable shift in Netflix's focus over the years, with a higher count of movies compared to TV shows in recent years. This is evidenced by the substantial number of movie releases, reaching 5,907, compared to 2,649 TV shows. Furthermore, the analysis identifies that the best times to launch content are July and December for movies, while December is optimal for TV shows, aligning with seasonal viewing patterns.

### **Univariate/Bivariate Plot Comments**

Univariate plots illustrate the distribution of attributes like genre count and duration, with boxplots effectively showcasing the differences in duration magnitude between TV shows and movies. The average TV show spans 1 to 2 seasons, while movies average around 100 minutes. Bivariate plots, including heatmaps, depict the relationship between countries and genres, revealing that the US and India dominate the content landscape with the highest numbers of Dramas, Comedies, and International Movies. This analytical approach aids in visualizing complex relationships and guiding content recommendations based on audience preferences and regional trends.

## **Business Insights**

### **Patterns Observed**

The analysis reveals several key patterns in Netflix's content offerings. Firstly, there is a consistent increase in the number of movie releases compared to TV shows over recent years, with a peak in 2019. This trend indicates a growing emphasis on movie content, especially during high-demand months like July and December. Additionally, the data shows that the top genres—Dramas, International Movies, and Comedies—are well-represented across key markets such as the USA and India, which also have the highest counts of cast and directors.

### **Implications**

These patterns have significant implications for Netflix's content strategy. The shift towards more movie releases suggests that Netflix is capitalizing on audience preferences for films, especially during peak viewing seasons. Understanding regional trends, such as the dominance of specific genres in different countries, can inform targeted marketing and content creation strategies. Furthermore, the high percentage of "Unknown" directors indicates potential gaps in data that could be leveraged for better recommendations and personalization efforts. By enhancing content offerings in genres that resonate most with viewers and refining data collection methods, Netflix can improve user engagement and retention.

## **Recommendations and Conclusion**

### **1. Increase Focus on Movies:**

Given the recent trend of higher movie releases compared to TV shows, Netflix should continue to prioritize movie production, particularly in high-demand months like July and December. This could involve increasing investment in original films and acquiring more exclusive content.

### **2. Diversify Genre Offerings:**

As dramas, international movies, and comedies are among the most popular genres, Netflix should consider expanding its offerings within these categories. This could include producing content that blends elements from multiple genres to attract a broader audience.

### **3. Leverage Regional Insights:**

Understanding that specific genres are more popular in certain countries, Netflix can tailor its content strategy to align with regional preferences. This might involve localized marketing campaigns and partnerships with local filmmakers to create culturally relevant content.

### **4. Enhance Data Collection and Analysis:**

The significant number of “Unknown” directors indicates gaps in data collection. Netflix should enhance its data analytics capabilities to gather more comprehensive information about its content creators. This will facilitate better recommendations and improve viewer engagement.

### **5. Invest in Shorter Formats:**

With an average TV show duration of 1 to 2 seasons, Netflix could explore shorter series formats or limited series that provide concise storytelling, catering to viewers who prefer quick consumption of content.

### **6. Monitor Global Trends:**

Regularly analyze global viewing trends and adjust content strategy accordingly. Understanding emerging markets and shifting viewer preferences can help Netflix stay ahead of competitors and maintain its leadership in the streaming industry.

**By implementing these recommendations, Netflix can strengthen its content strategy, enhance user experience, and continue to grow its subscriber base effectively.**